Fractal Clustering

Session 3

Dr. Ali Arsanjani

@AliArsanjani

Adjunct At San Jose State University

Senior Lecturer UCSD

Sanders wins New Hampshire

New Hampshire Primary

CANDIDATE	VOTES	%
SANDERS	71,410	26.0%
BUTTIGIEG	67,044	24.4%
KLOBUCHAR	54,244	19.7%
WARREN	25,612	9.3%

MORE CANDIDATES›

est 94% in

updated 11:55 PM ET, Feb 11, 2020

All 4 federal prosecutors quit Stone case after DOJ pushes to reduce Trump ally's sentencing

Trump withdraws Treasury nomination of ex-US attorney who oversaw Stone prosecution

Analyst on Stone: This stinks to high hell

Opinion: This should disqualify Michael Bloomberg from the 2020 race

Bennet drops out after after failing to gain

Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data.

- Sales and Marketing organizations are hiring and are structured to deal with these current clusters
- Accounting an finance position their work on this basis
- Sales projections and company strategy is dependent on these

Low market

mid market

top market

Small Clients

Medium Clients

Enterprise Cllients

Motivation

https://medium.com/uptick-blog/stock-picks-using-k-means-clustering-4330c6c4e8de

Stock Picks using K-Means Clustering

Timothy Ong Follow

Feb 8, 2019 · 6 min read

Disclaimer: I am not vested in any of these stocks and I am not an equity analyst. Please do your due diligence before investing!

TLDR: Wanted to pick the best stocks to invest. Used K-means clustering to filter out a winning group. Discovered a group of 57 stocks with outstanding performance.

From the motivating stock purchase use case:

- “narrow down my search for the winning stocks.
- I plan to use a clustering technique by grouping similar stocks,
- and hopefully be able to filter out the better performing stocks.”

The data set that I have obtained was pulled using the Stocker and Yahoo-Finance python packages. Below is an example of how the data set looks like.

ticker	date	year	open	close
MCBC	2012-01-03	2012	2.28	2.26
MCBC	2012-01-04	2012	2.22	2.26
MCBC	2012-01-05	2012	2.26	2.24
MCBC	2012-01-06	2012	2.24	2.27
MCBC	2012-01-09	2012	2.20	2.25

Example of how the raw data looks like

Objective Functions create new Computed Features:

Based on the data set, I will fit these two variables into the K-means model:

1. Annual returns
2. Annual variance

stocks where the average annual returns was 24% with a variance of 5%

I have decided to use these variables as they inform us on the stock performance and its volatility (risk).

add these to your data set or cerate a new data set based on them

From the raw data, I have transformed it into a more usable data frame that informs me of each stocks average annual return and variance (over the last 7 years).

ticker	avg_yearly_returns	yearly_variance
AAN	0.080143	0.048170
AAON	0.100059	0.079516
AAP	0.165665	0.124715
AAPL	-0.005033	0.161882
ABC	0.131880	0.083416

Example of how the transformed data looks like

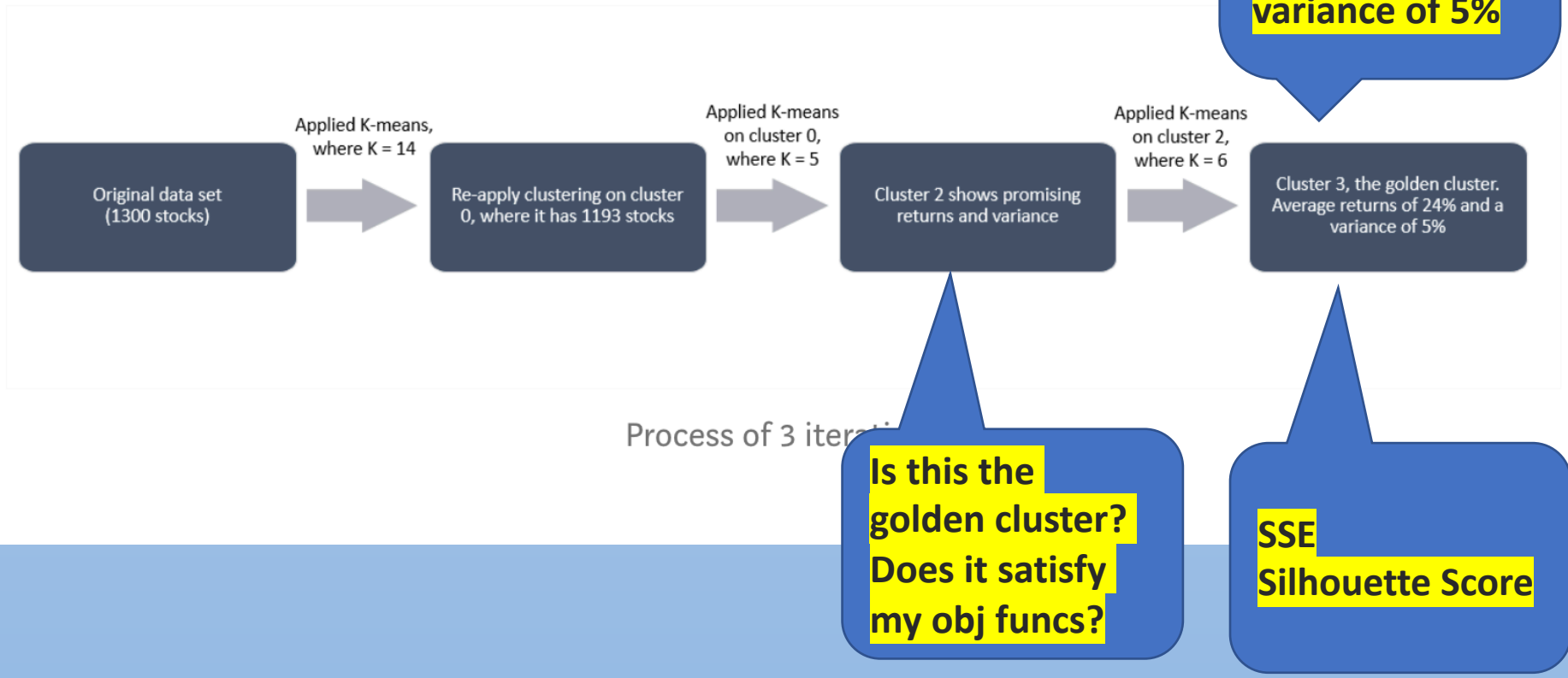
Metrics: measure the results you obtained in your objective functions for the new features

Evaluating the Model (K-means)

I have used two metrics to evaluate the model:

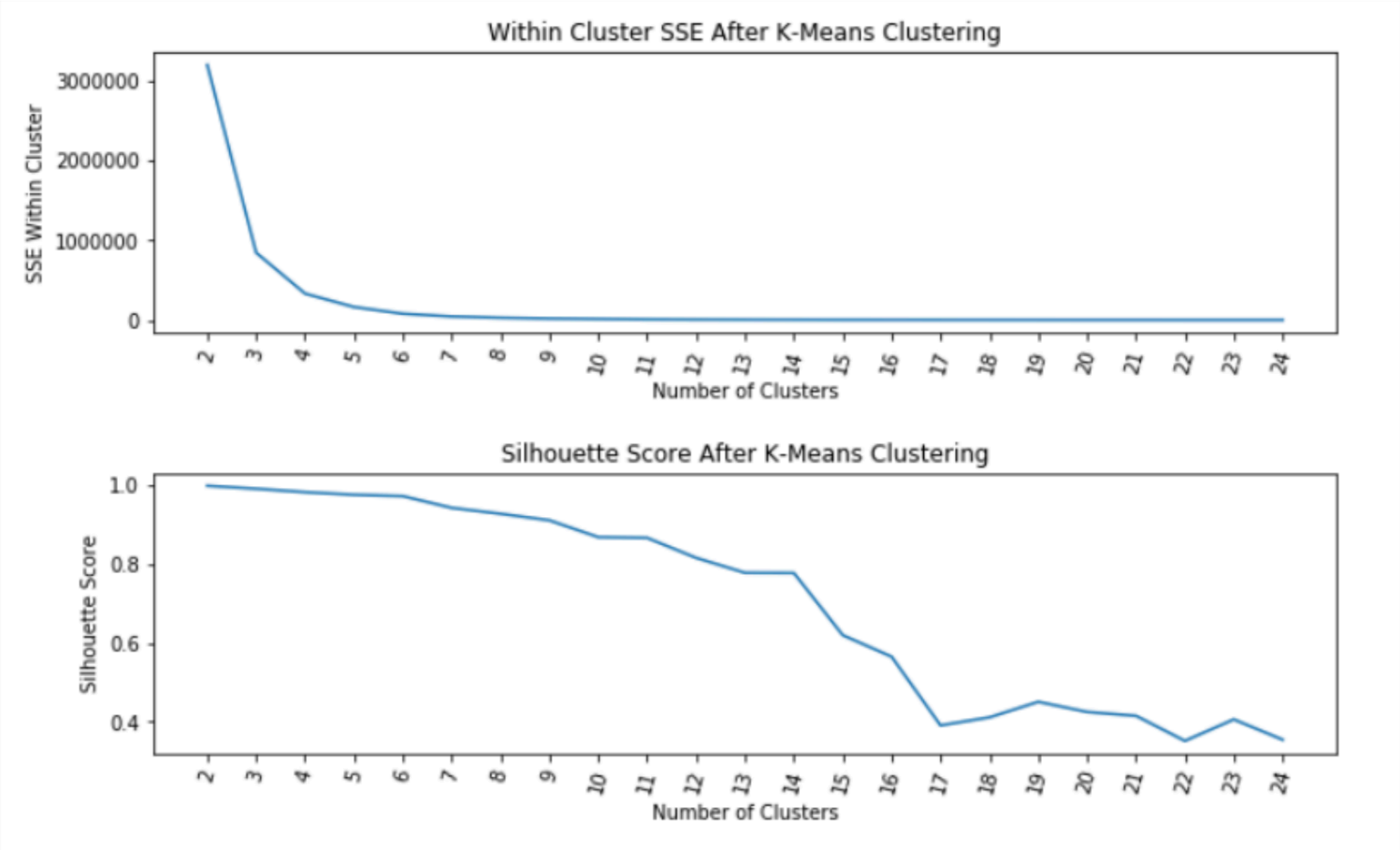
1. Sum of squares of error (SSE) within cluster. SSE value will inform the user on how close each data points are to the center.
2. Silhouette score. Silhouette score measures how similar the data point is to its own cluster compared to other clusters.

Apply Fractal Clustering



<other features>	Cluster	Iteration	Obj func. returns	Obj func. var	SSE	Silhouette
	1	1				
	1.0	2				
	1.0.2	3				

For example, in my first iteration I have looped the value of K 25 times, from K=1 to K=25. This resulted in the two graphs seen below.



Deciding my K value using these 2 graphs

cluster	avg_yearly_returns	yearly_variance	ticker
0	0.092419	0.110761	1193
11	0.336065	1.018290	59
8	0.593508	2.854767	19
12	0.837786	5.224438	11
6	1.067458	8.362185	5
10	1.331071	15.027225	3
4	1.836886	25.347059	3
9	2.420434	43.899749	1
13	2.602269	30.862794	1
7	3.340888	57.203435	1
3	3.562943	72.620209	1
5	3.857160	112.449844	1
2	5.259784	187.519424	1
1	9.074862	651.720274	1

Clustering result of the 1st iteration

Second Iteration Resulted in 2 Potential Clusters

In the second iteration of K-means clustering, I have chosen my K value to be 5 (using the same logic as above). The results (as seen below) are better spread out and we can see that cluster 0 and cluster 2 are the better performing clusters.

cluster	avg_yearly_returns	yearly_variance	sharpe_ratio	ticker
0	0.081021	0.053163	0.342037	603
1	0.297600	0.368385	0.480800	65
2	0.206852	0.115717	0.663048	257
3	0.044477	0.303681	0.059671	84
4	-0.080652	0.113519	-0.309757	184

Clustering results of the 2nd iteration

Find a Domain specific Metric to measure your objective functions

It was quite difficult discerned which was the better performing cluster (out of the two) as the average annual return and variance were quite proportionate. As such, I decided to add a metric, **Sharpe Ratio**, that better reflects the stock performance.

Sharpe Ratio is a measure that helps us understand the return of an investment compared to its risk. You can read more about it here if you are interested.

To have a better visualization of the numbers I have plotted out 2 box-plots.

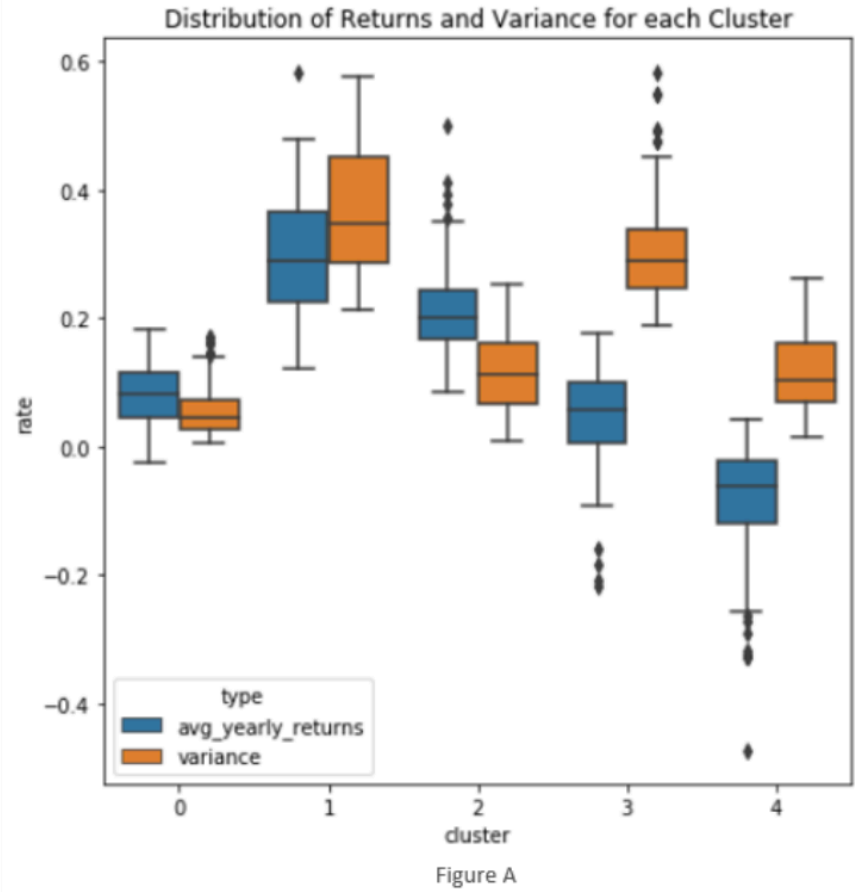


Figure A

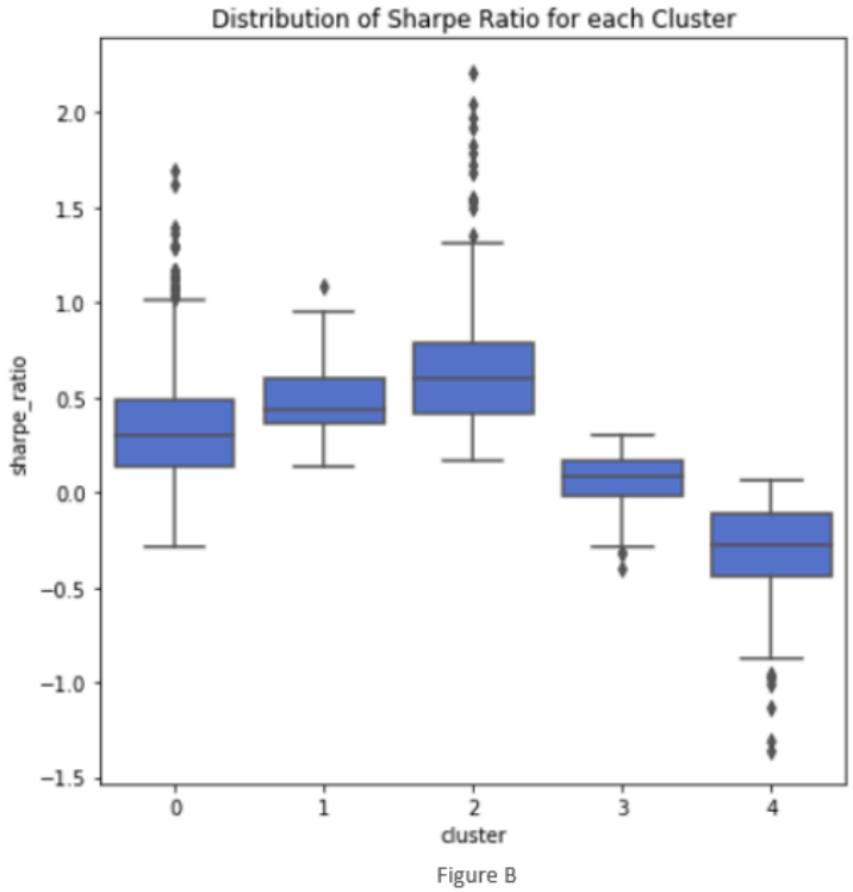


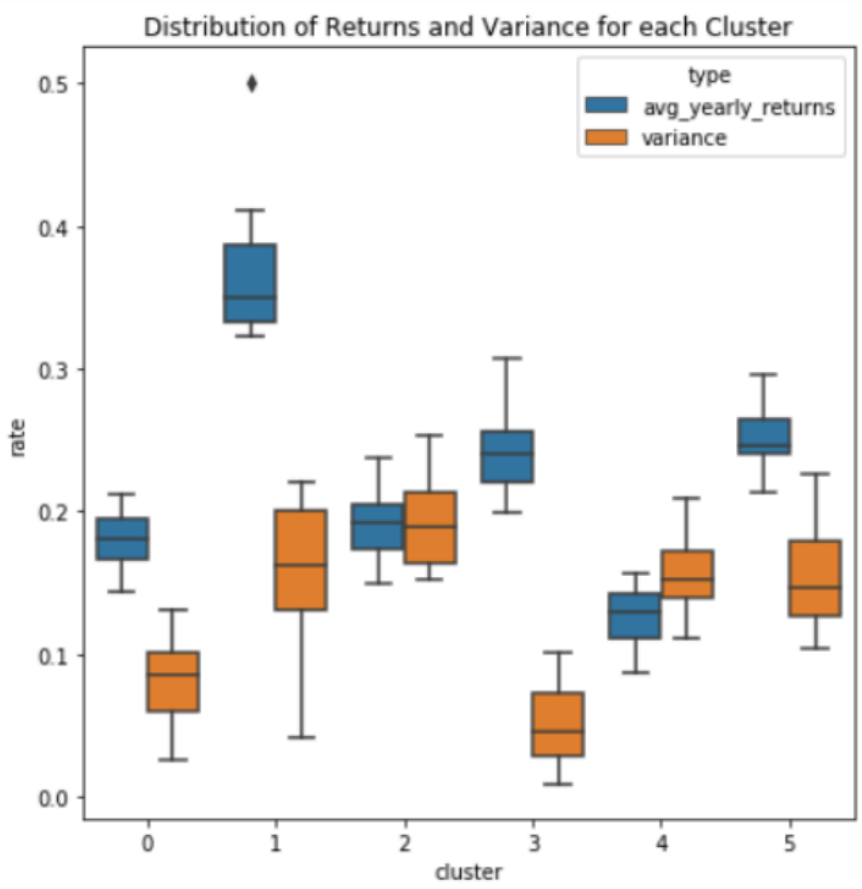
Figure B

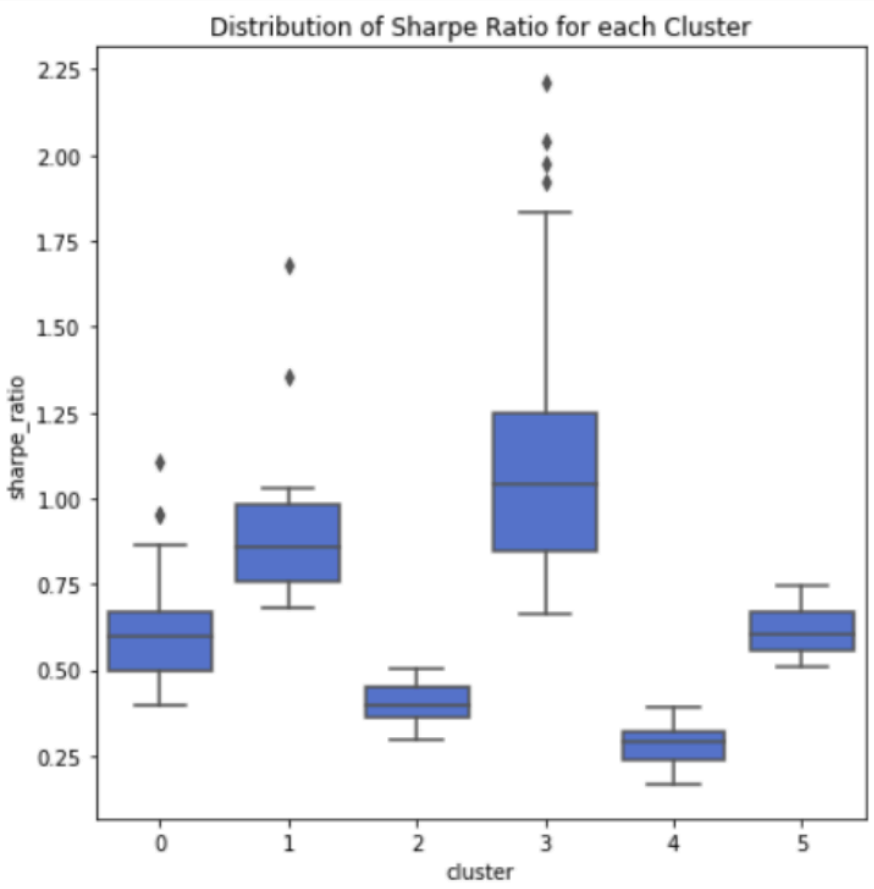
Third Iteration Resulted in the Golden Cluster

In the third iteration of K-means clustering, we can find the golden cluster! Cluster 3 has an average annual returns of 24%, a variance of 5% and its Sharpe Ratio ranged from 0.7 to 2.2!

cluster	avg_yearly_returns	yearly_variance	sharpe_ratio	ticker
0	0.081021	0.053163	0.342037	603
1	0.297600	0.368385	0.480800	65
2	0.206852	0.115717	0.663048	257
3	0.044477	0.303681	0.059671	84
4	-0.080652	0.113519	-0.309757	184

Golden cluster found in third iteration





Cluster 3 is our clear winner

GitHub

- https://github.com/ttimong/blog-posts/blob/master/blog1-kmeans-clustering/final_model.ipynb

Triangulate Feature Types/ categories into Clusters

- Choose a set of feature types or categories and combing them into a cluster you expect to see

Fractal Clustering

1. Apply the appropriate clustering technique eg GMM as a macro cluster
    1. Apply Agglomerative if data is hierarchical
2. Apply K-means recursively in each of the GMM clusters
    1. Stop when standard deviation > 3 standard deviations
3. Note in small groups, the data will likely be convex so you can apply k-means, if not continue with the best fit algorithm

Exercise

- For your designated dataset [which you can change or add to!]
- Decide on objective function
  - E.g., annual return, annual variance
  - What additional features should I add to my existing dataset?
  - Compute the objective function for each cluster, add them to your data set
- Measure them:
  - Metrics
- Visualize
- Silhouette Scores
- Elbow
- Apply Fractal K-means 2-3 [3-5] times

Exercise: Seekers Housing Prices

- For your designated dataset [which you can change or add to!]
- Decide on objective function
  - E.g., annual return, annual variance
  - What additional features should I add to my existing dataset?
  - Compute the objective function for each cluster, add them to your data set
- Visualize
- Measure them: apply metrics to see if you have met the objective functions
  - Metrics: Sharpe Ratio
  - Silhouette Scores
  - Elbow
- Apply Fractal K-means 2-3 [3-5] times

- Find a cluster of houses to invest in.

Exercise: Movie Recommendation – Team Sigma

- For your designated dataset [which you can change or add to!]
  - IMDB
- Decide on objective function
  - E.g., maximize? Minimize?
    - Maximize number of viewing/viewers per movie or recommend to others (thumbs up)
      - Ratings
      - Viewings → PCA something that correlates to
  - What additional features should I add to my existing dataset?
  - Ranking for genre
  - Compute the objective function for each cluster, add them to your data set
- Visualize

- Measure them: apply metrics to see if you have met the objective functions
  - Metrics: research metrics on movies and recommendations
  - What are the factors that cause a series to be cancelled?
    - Predict which series will be cancelled and which will be renewed.
  - Silhouette Scores
  - Elbow
- Apply Fractal K-means 2-3 [3-5] times

- Find a cluster of houses to invest in.

Exercise: Feature Finders Bank Loan Approval Prediction

- For your designated dataset [which you can change or add to!]
- Look for bias in the loan approval process

- Decide on objective function
  - E.g., decrease biased loan verdicts

  - What additional features should I add to my existing dataset?
  - Compute the objective function for each cluster, add them to your data set
- Measure them:
  - Metrics:
    - Balance in the dataset, bias in the data, labels , etc.
    - Cluster A : loan approved 75% of the time based on various bias factors (figure out)
      - Ethnicity, income ,location , credit score, num of deps
    - Cluster B: loan approved 35% of the time
- Visualize
- Silhouette Scores
- Elbow
- Apply Fractal K-means 2-3 [3-5] times

Exercise: Shining Unicorns Prediction

- For your designated dataset [which you can change or add to!]
  - Age of persons , male/female, spending score
- Decide on objective function
  - E.g., age bracket of who is more likely to spend
  - Malls / brand store target age groups ; age and spending correlation, increase overall profit of the mall based on increased spending from more age groups, find current age group spending, and incentivize them to spend more!
  - What additional features should I add to my existing dataset?
  - Compute the objective function for each cluster, add them to your data set
- Measure them:
  - Metrics: increase spending
- Visualize
- Silhouette Scores
- Elbow
- Apply Fractal K-means 2-3 [3-5] times

Exercise: Cereal Killers – Grocery store recommendation

- For your designated dataset [which you can change or add to!]
  - Age of persons , male/female, spending score
- Decide on objective function
  - know your customers!!! → segmentation
  - Clustering for hyper-personalization
  - "how low can you go!!" how small can your clusters get! Target the market of one
  - Political elections are swayed through hp!
  - Brexit →
  - What additional features should I add to my existing dataset?
  - Compute the objective function for each cluster, add them to your data set
- Measure them:
  - Metrics: increase spending
- Visualize
- Silhouette Scores
- Elbow
- Apply Fractal K-means 2-3 [3-5] times

Exercise: Go ML– Covid-19 Dataset

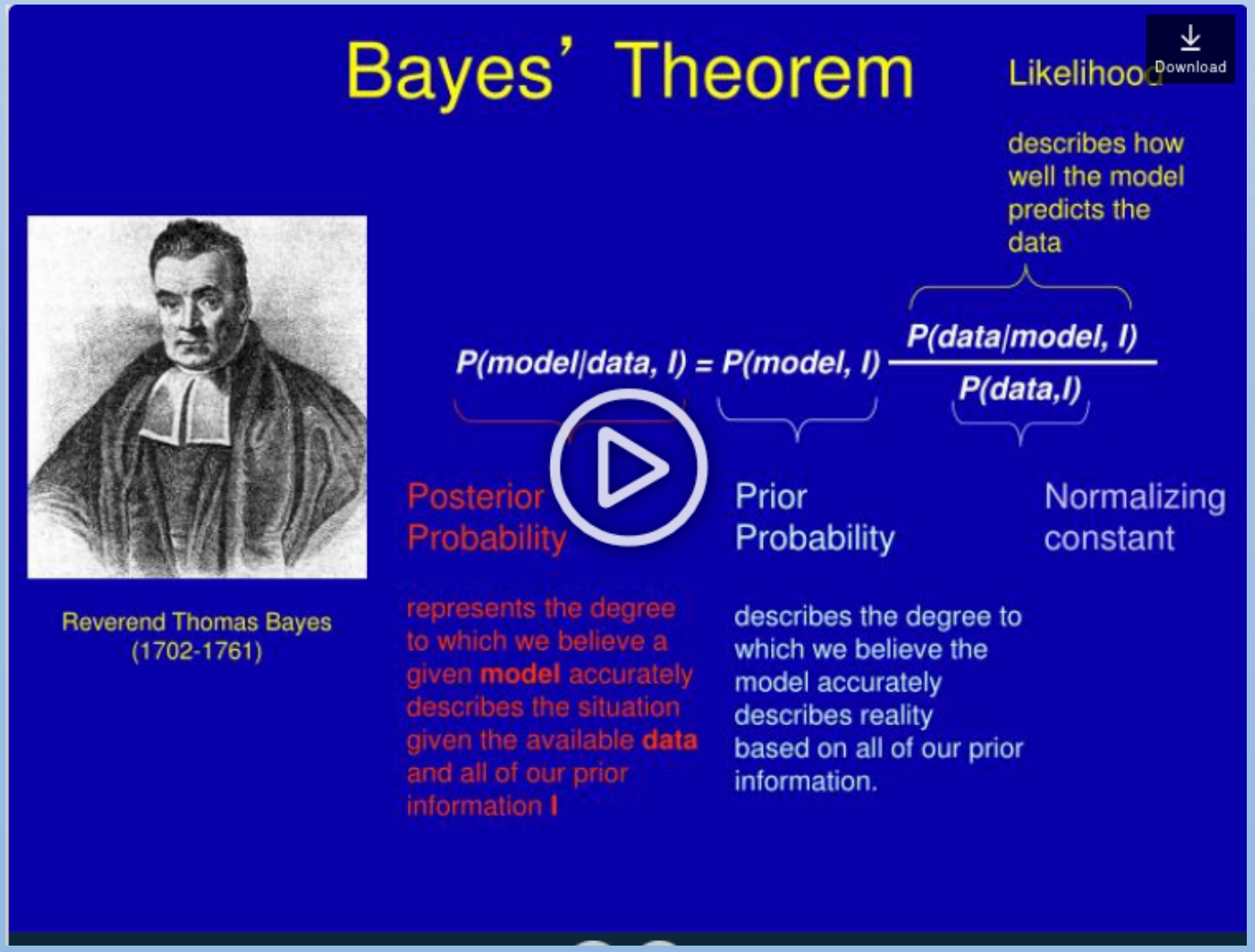
- For your designated dataset [which you can change or add to!]
  - County datasets
- Decide on objective function
  - recommend counties where they should be focus on separating two groups: contracted and not contracted
    - social separation and increased testing
    - Segmentation between likely, known and possible infections / spread
    - Asymptomatic, symptomatic, still healthy
  - What additional features should I add to my existing dataset?
  - Compute the objective function for each cluster, add them to your data set
- Measure them:
  - Metrics: spread
    - Low, med, hi, critical
- Visualize
- Silhouette Scores
- Elbow
- Apply Fractal K-means 2-3 [3-5] times

Exercise: Blast-off – patient data

- For your designated dataset [which you can change or add to!]
  - Age group
- Decide on objective function
  - likelihood of cardiac disease in younger age groups in order to prevent card dis or to avert risk factors of card anomalies
  - What additional features should I add to my existing dataset?
  - Compute the objective function for each cluster, add them to your data set
  - Compute the bayes theorem as your objective function
  - Cluster on highest risk factors
  - Cluster on age and compare to higher risk factor clusters!
    - Given the data on the small cluster or person today, what is the likelihood they may contract some disease
    - Golden cluster == highest risk cluster
  - Obj : probability or probability distribution
- Measure them:
  - Metrics: symptomatic vs asymptomatic
  - Likelihood of
    - Prediction of when they might contract if they continue to consume 7000 calories a day!!!
    - Sit all day [daily habits, exercise, activity]

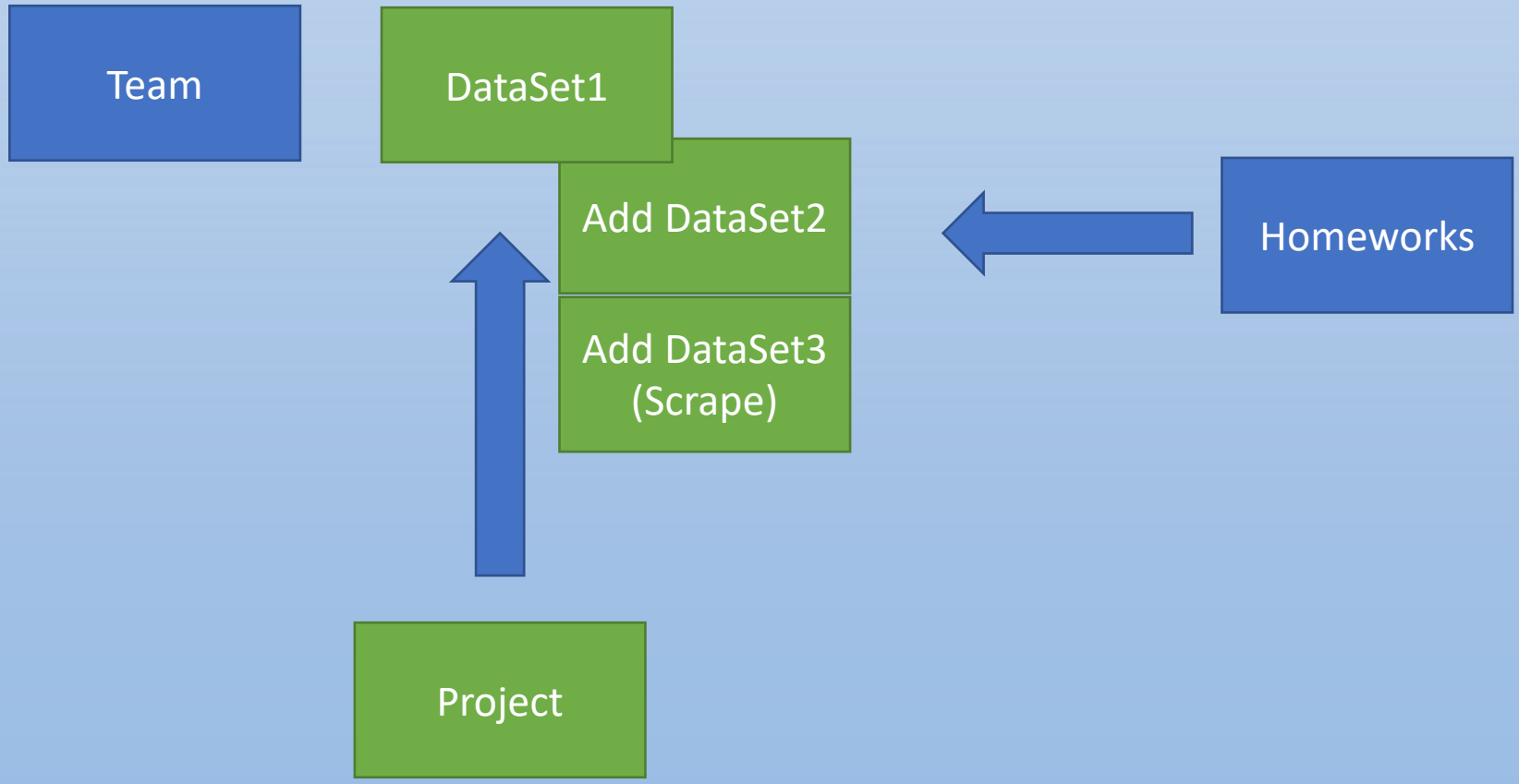
- Visualize
- Silhouette Scores
- Elbow
- Apply Fractal K-means 2-3 [3-5] times

Maximum Likelihood, Bayes Theorem



Final outcome is finding the Golden Cluster!!!

- Customer segment that is you should recommend groceries to
- How should I identify the GC in my domain??
- Recomm you make to your golden cluster will cause
  - Purchases to go up
  - Increase mall sales
  - Most likely to have/will contracted/contract


Team
DataSet1
Add DataSet2
Add DataSet3
(Scrape)
Homeworks
Project

Similarity during amalgamation

- Cosine


Deep
Context
AI